国家超级计算天津中心

TH-1A 大系统用户手册

二〇一三年四月

目 录

1 TH-1A 大系统运行环境 1
1.1 平台架构 1
1.1.1 登陆节点 1
1.1.2 计算节点 2
1.1.3 存储节点 2
1.1.4 数据拷贝节点 3
1.1.5 管理节点 3
1.1.6 互联网络 3
1.2 系统环境 3
1.3 编译环境 4
1.3.1 Intel 编译器 4
1.3.2 gcc 编译器 6
1.3.3 mpi 编译环境 6
1.3.4 CUDA 编译环境 7
1.3.5 其它环境（Python 等） 8
1.4 软件环境 8
1.4.1 IDL 软件 8
2 TH-1A 大系统访问方式 10
2.1 基本条件 10
2.2 登陆 VPN 10
2.2.1 Windows 系统 10
2.2.2 Linux 系统 14
2.2.3 Mac 系统 17
2.2.4 VPN 登陆注意事项 19
2.3 登陆服务器和数据传输 20
2.3.1 登陆服务器 20
2.3.2 文件传输 22
2.4 环境变量设置 23
2.5 退出系统 24
2.6 用户帐号密码修改 24
3 作业提交 25
3.1 使用限制 25
3.1.1 分区限制 25
3.1.2 用户限制 26
3.1.3 磁盘配额限制 27
3.2 状态查看命令 28
3.2.1 节点状态查看 yhinfo 或 yhi 28
3.2.2 作业状态信息查看 yhqueue 28
3.3 提交作业 29
3.3.1 批处理作业 yhbatch 30
3.3.2 交互式作业提交 yhrun 33
3.3.3 分配模式作业 yhalloc 35
3.4 任务取消 yhcancel 36
4 常见问题 38
4.1 VPN 登陆问题 38
4.2 系统登陆问题 39

4.3 作业运行问题 39
4.4 存储问题 41
5 技术支持 42
附录 A 常用 Unix 命令 43
A1 基本命令 43
A2 目录操作 43
A3 文件创建、复制与删除 43
A4 文件属性 43
A5 文件显示与连接 44
A6 文件查找与比较 44
A7 文件压缩与备份 44
A8 输入输出重定向 44
附录 B 常用 vi 命令 45
B1 进入与退出 vi 45
B2 移动光标 45
B3 正文输入、删除、替换、恢复和查找命令 45
B4 行编辑命令 46
附录 C GDB 常用命令 47
C1 启动 gdb 47
退出 gdb 47
执行程序 47
程序执行控制 47
断点和观察点 47
程序栈帧 48
数据显示 48

TH-1A 大系统用户手册

1 TH-1A 大系统运行环境

1.1 平台架构

TH-1A 大系统由登陆节点、计算节点、存储节点、管理节点、数据拷贝节点以及天河高速互联网络组成。其平台架构如下图 1.1 所示：

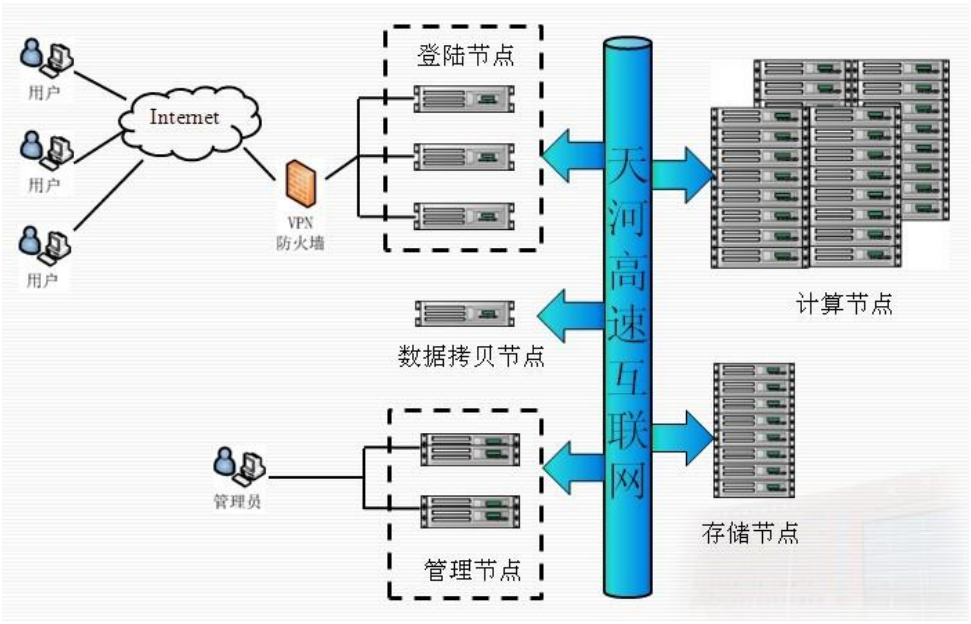

图 1.1 系统平台架构

对 TH-1A 大系统平台架构以及功能说明如下。

1.1.1 登陆节点

登陆节点采用完整安装的麒麟操作系统，挂载共享存储，其上具备软件编译所需的编译环境。目前对用户开放的登陆节点有 LN0-LN3，**其中 LN0 作为 MAC 用户的登陆和数据传输节点，具体登陆方式见“2.2 登陆和数据传输”中的 MAC 系统登陆部分；LN1-LN2 作为 windows 与 linux 用户的登陆节点，LN3 作为**

windows 与 linux 用户的数据传输节点，具体登陆方式见“2.2 登陆和数据传输”中的 windows、linux 系统登陆部分。

作用：登陆节点为用户提供一个登陆系统的平台，用户可以通过 internet 网络登录 VPN，然后通过 ssh 终端登录到登陆节点上。

允许操作：在登陆节点上用户可以进行软件编译与调试，环境变量配置，作业提交，文件编辑，结果查看等操作。

禁止操作：禁止用户直接在登陆节点上运行计算程序。

1.1.2 计算节点

计算节点本身没有本地硬盘，采用 ramdisk 精简内核系统，挂载共享存储，具备软件运行所需的运行环境。

计算节点采用 CPU+GPU 的架构，其具体配置如表 1 所示：

CPU*2	型号	Intel Xeon X5670	核心数量	6	主频：	2.93GHz
GPU	型号	Fermi M2050	核心数量	448	显存：	3GB
内存	24 GB					
硬盘	无					
操作系统	版本	RHEL 5.3	内核	2.6.32.16-TH		

作用：计算节点为用户提供一个大规模并行计算资源，用户可以将自己的作业通过作业调度系统提交到计算节点上运行。计算节点上具备程序运行所需的运行环境，但不具备软件编译环境。

用户在无作业的情况下无法直接登录到计算节点上，但可以通过 ssh 服务登录到正在运行用户自身作业的计算节点上查看自己程序的运行情况。

1.1.3 存储节点

TH-1A 大系统采用分布式存储文件系统，该文件系统由多个存储节点构成，对外提供一个统一的大分区，供所有登陆节点与计算节点进行挂载。

作用：存储节点用于存储用户的所有数据文件，以共享方式使登陆节点与计算节点都可以对用户数据进行读写操作。

为了满足并行计算对共享存储的高速读写需求，分布式存储不提供多副本服

务，因此无法长期保证用户数据的安全，建议用户及时拷贝自己的核心数据结果。中心为了保障绝大部分用户的使用体验，对用户存储空间进行了限制，详见“3.1 小节的磁盘配额限制”，希望大家及时清除共享存储上的无用数据。

1.1.4 数据拷贝节点

数据拷贝节点为 NAS 存储服务器，**支持 EXT3 文件系统**，挂载共享存储，可直接通过 SATA 硬盘为用户提供数据拷贝服务。

作用：用户可以将 SATA 硬盘直接邮寄至超算中心，并通过邮件联系应用部负责人员，说明需要拷贝数据的目录，由超算中心管理员通过数据拷贝节点将共享存储上的数据拷贝到您的 SATA 硬盘上，再由应用部负责人员将硬盘邮寄给您。

1.1.5 管理节点

管理节点为用户提供用户登录认证，作业调度等服务。

作用：管理员通过管理节点对用户认证，作业调度等服务进行管理，对系统状态进行监控、对节点问题进行处理。管理节点禁止用户访问。

1.1.6 互联网络

TH-1A 大系统的互联网络由天河高速互联网络构成，这是一种高性能通信互联技术，具有超高通信效率，超低通信延迟的特点。

作用：在天河系统中天河高速互联网络主要用于支持并行任务间的通信，并实现全局文件系统的数据传输。

1.2 系统环境

- 共享目录：/vol-th

  该目录下的全部文件在所有登陆节点与计算节点上都可以访问。

- 用户根目录：/vol-th/home/

  该目录属于共享目录的下级目录，用于存放用户的数据，在创建用户时会

在该目录下产生一个与用户名相同的目录，用户每次登陆系统后会自动跳转到该目录下，用户的数据及文件默认存储在与自己同名的目录中。

- 常用软件安装目录：/vol-th/software

  该目录属于共享目录的下级目录，存放用户常用的软件或编译器等。该目录仅提供大众用户普遍需要使用的常用软件，会随着用户的需求不断更新，软件由管理员负责安装，为用户提供使用。

- 常用动态链接库目录：/vol-th/lib

  该目录属于共享目录的下级目录，存放可执行程序运行时调用的一些系统以及编译器的动态链接库，供用户使用，其中包括 MKL 库/vol-th/lib/mklem64t 等。

1.3 编译环境

在 TH-1A 大系统的登陆节点中，目前安装了 Intel 编译器和 GCC 编译器。用户可根据自身程序需求，选择相应的编译器进行编译和应用程序开发，**由于 TH-1A 大系统广泛采用了 Intel 的 CPU，因此在编译中除特定需要，建议用户首选 Intel 编译器**。另外，在 TH-1A 大系统的登陆节点上还提供了 MPI 并行编译环境，以及针对 GPU 的 CUDA 编译环境。下面将分别具体介绍各编译器及编译环境。

1.3.1 Intel 编译器

TH-1A 大系统上安装了 Intel 编译器，版本为 11.1，支持 C，C++，Fortran77 和 Fortran 90 语言程序的开发。Intel 编译器安装在/opt/intel/Compiler/11.1/059 目录，其中：C 和 C++编译器，以及 Fortran 77/90 的相应命令程序均在：/opt/intel/Compiler/11.1/059/bin/intel64/中，编译命令分别为 icc 和 icpc，ifort 等。

用户在登陆节点上使用 Intel 编译器进行程序编译时需添加如下环境变量声明：

```
source /opt/intel/Compiler/11.1/059/bin/intel64/iccvars_intel64.sh
source /opt/intel/Compiler/11.1/059/bin/intel64/ifortvars_intel64.sh
```

用户在计算节点上提交作业运行时，如需要调用 Intel 编译器的动态库，则需要添加如下环境变量声明：

```
export LD_LIBRARY_PATH=/vol-th/lib:$LD_LIBRARY_PATH
```

Intel 11.1 对应的 mkl 安装路径为/opt/intel/Compiler/11.1/059/mkl，用户可以使用该目录下的 lib/em64t 的 mkl 库。

用户在使用 mkl 库进行程序编译时需添加如下环境变量声明：

```
source /opt/intel/Compiler/11.1/059/mkl/tools/environment/mklvarsem64t.sh
```

用户在在计算节点上提交作业运行时，如需要调用 mkl 动态库，则需要添加如下环境变量声明：

```
export LD_LIBRARY_PATH=/vol-th/lib/mklem64t/:$LD_LIBRARY_PATH
```

常用编译选项如下：

（1） 优化选项

-O0：禁止优化

-O1：优化代码大小和代码局部性。

-O2（缺省值）：优化代码速度（推荐使用）

-O3：-O2+激进的优化（循环、存储访问转换、预取）。需要注意的是，-O3 并不一定适合所有程序。

-fast：打开-O3、-ipo、-static、-no-prec-div 和 －xP

-ipo：过程间优化

（2） 输出和调试选项

-c：只生成目标文件

-S：只生成汇编文件

-g：调试选项

-o <file>：指定生成的输出文件名

（3） 浮点选项

-mp：维持浮点精度（禁止某些优化）

-mp1：改善浮点精度。和-mp 相比，-mp1 对性能影响较小

（4） 链接选项

-L<dir>：指定链接时搜索的库路径

-l<string>：链接特定库

-static：静态链接

-shared：生成共享库

1.3.2 gcc 编译器

TH-1A 大系统上操作系统自带的 GCC 版本是 4.1.2，相关的编译命令都在/usr/bin 目录中。

为了满足用户对 GCC 更高版本的需求，目前 TH-1A 大系统安装了 GCC 4.6.1 版本，用于支持用户编译使用。安装路径为/vol-th/software/gcc-4.6.1，用户使用时需要添加如下环境变量：

```
export PATH=/vol-th/software/gcc-4.6.1/bin:$PATH
export LD_LIBRARY_PATH=/vol-th/software/gcc-4.6.1/lib64:$LD_LIBRARY_PATH
```

后续我们还会根据用户的需求在/vol-th/software 目录对 GCC 版本进行更新。

1.3.3 mpi 编译环境

由于 TH-1A 大系统包括两种基本编译环境，Intel 编译器和 gcc 编译环境，因此为了适应用户需要，系统的 mpi 编译环境包括两部分，即底层分别用 Intel 编译器和 gcc 编译器编译的 mpi 版本。由于 TH-1A 采用了自主互连的高速网络，因此底层 mpi 为自主实现，分别基于 Intel 和 GCC 编译器进行编译。**用户使用天河系统提供的 MPI 进行并行编译可以充分发挥天河高速网的性能，提高并行效率，这里给用户推荐使用基于 intel 编译器编译的 mpi（如 mpi_1.2.1_intel_11.1，mpi_1.4.1_intel_11.1）。**但如果用户的程序有特定的 mpi 版本需求，用户也可以在自己的根目录下安装所需要的 mpi。

天河系统提供的 mpi 编译器在/vol-th/software/mpi/目录下，目录命名格式为 mpi_aaa_BBB_ccc_ddd，其中 aaa 为使用的 mpi 版本，BBB 为底层采用的 Intel 编译器或者 GCC 编译器，ccc 为相应编译器的版本，ddd 为编译选项标注。后续我们会针对编译器的版本更新同时更新 mpi 的版本，请用户关注此目录的 mpi 版本。

用户使用 mpi 编译器进行程序编译时需添加如下环境变量声明：（以 mpi_1.2.1_intel_11.1 为例）

```
source /opt/intel/Compiler/11.1/059/bin/intel64/iccvars_intel64.sh
source /opt/intel/Compiler/11.1/059/bin/intel64/ifortvars_intel64.sh
export PATH=/vol-th/software/mpi/mpi_1.2.1_intel_11.1/bin:$PATH
export LD_LIBRARY_PATH=/vol-th/software/mpi/mpi_1.2.1_intel_11.1/lib:$LD_LIBRARY_
    PATH
```

并行 mpi 编译环境使用注意事项：

1. TH-1A 大系统安装了两种底层编译环境的 mpi，程序如无特殊需要，推荐使用/vol-th/software/mpi/目录下的 mpi。且该 mpi 的库均为静态库，用户不用担心动态链接库问题；

2. 原手册中介绍的系统登陆节点上/usr/local 目录下的 mpi 仍然可用，但后续不再对该目录中的 mpi 进行更新；

3. TH-1A 大系统具备自主高速互联网络，并提供 MPI 编程环境，如用户必须使用其他版本 mpi，比如 openmpi1.4.8，mpich2-1.3.1 等，也可以自己安装并部署。用该 mpi 编译的程序，同样可以利用高速互联网络的虚拟以太网运算任务，但性能会较 TH-1A 自主 MPI 低很多。

MPI 编译命令内部会自动包含 MPI 标准头文件所在的路径，并自动连接所需的 MPI 通信接口库，所以不需要用户在命令行参数中指定。

如果用户使用 makefile 或 autoconf 编译 MPI 并行程序，还可以将 makefile 中的 CC，CXX，F77，F90 等变量设置成 mpicc，mpicxx，mpif77，mpif90，或这在 autoconf 的 configure 过程前设置 CC，CXX，F77 和 F90 等环境变量为 mpicc，mpicxx，mpif77 和 mpif90 等。

1.3.4 CUDA 编译环境

由于计算节点，每个节点包括一个 M2050 GPU，因此 LN0-LN3 中，有相应的 CUDA 编译环境。

CUDA 编译环境包含三个部分，编译器、SDK 和设备驱动，目前计算节点

CUDA 编译环境已经更新至 CUDA4.0，用户可以选择相应的编译器。

CUDA 编译器及 SDK 部署在/vol-th/software/cuda 目录下，请用户选择 cuda-4.0 使用。其中 cuda-4.0 为 CUDA 4.0 编译器；目前我们已经有了 CUDA4.1 的环境，其他环境也可以提供，但是由于驱动的原因需要重启节点，因此用户使用时需要提前联系我们。此外，该目录下还有包括 3.0，3.1 和 3.2 在内的三套早版本的 cuda 编译器。

用户使用 CUDA 进行程序编译时需添加如下环境变量声明：（以 cuda4.0 为例）

```
export PATH=/vol-th/software/cuda/cuda-4.0/bin:$PATH
export LD_LIBRARY_PATH=/vol-th/software/cuda/cuda-4.0/lib64:$LD_LIBRARY_PATH
```

注意：目前节点的 CUDA 版本已经升级至 4.0，因此请大家在编译运行 GPU 程序时，选择使用 CUDA4.0 的编译和运行环境，以及相应的动态链接库。原手册中提供的一些 CUDA 编译运行环境仍有效，但后续不再对其进行更新。

1.3.5 其它环境（Python 等）

目前 TH-1A 大系统还安装了诸如 Python 等运行环境，python 版本为 2.7，安装目录为/vol-th/software/python2.7，用户使用时可以进行选择，通过设置相应的环境变量如下：

```
export PATH=/vol-th/sottware/python2.7/bin:$PATH
```

1.4 软件环境

1.4.1 IDL 软件

目前 TH-1A 系统 LN2 节点安装了版本为 8.2.1 的 IDL，包含主模块 IDL 8.2-CON 和扩展模块 IDL 8.2-ADVANCED-CON，该版本支持 32 位和 64 位系统。

IDL 软件安装目录：/opt/exelis

ENVI 启动命令：

```
envi                启动 ENVI+IDL
```

envi_r	启动 ENVI
envihelp	启动 ENVI 帮助
envi –classic	启动 ENVI Classic 版本
exelislicense	启动 ENVI 许可管理

IDL 启动命令：

idl	启动 IDL 命令行开发环境（默认启动 64 位的 IDL）
idlde	启动 IDL 图形开发环境（默认启动 64 位的 IDL）
idl –vm	启动 IDL 虚拟机
idl –rt	启动 IDL Runtime
idldemo	启动 IDL Demo，同样可以在 IDL 命令行输入 demo 启动
idlhelp	启动 IDL 帮助
idl -32	启动 32 位 IDL 命令行开发环境

注意事项：

1. 因安装该软件时绑定到 LN2 登陆节点的物理地址，目前只可在 LN2 上使用该软件。

2. 因购买的 license 限制，在同一时刻仅允许一个用户使用该软件。

3. 在天河系统下，若需要在 idl 图形界面开发环境下使用 ADVANCED-CON 时，使用命令 idlde -outofprocess 启动 idl 图形开发环境，这样在该图形开发环境下即可支持 ADVANCED-CON。默认不使用 outofprocess 时，IDL 使用 Java Windows 来显示图形，使用该选项时，IDL 使用 X11 窗口显示图形。

2 TH-1A 大系统访问方式

为了更好的保证用户的数据安全，中心采用 SSL VPN 实现远程用户对天河系统的登陆访问，用户需要首先登陆 VPN 后，才能使用中心资源。下面将详细介绍用户如何登陆 TH-1A 大系统。

2.1 基本条件

用户需要具备的基本条件如下：

1. 经过了中心用户基本审查创建流程，并填写了**相应的文件和协议**。

2. 具备一个 **VPN 账号及密码**。（用户如经过了审查创建流程，会收到用户账号创建成功的 email，里面会有 VPN 及系统的账号及密码）

3. 具备一个**系统用户及密码**。

具备上了上述条件，您就可以尝试登陆至 TH-1A 大系统使用系统资源。首先登陆 VPN，然后登陆服务器，步骤及所需软件如下节描述。

2.2 登陆 VPN

目前 TH-1A 系统已经同时接入了联通和电信双网络，用户根据自身网络接入商的不同可以选择不同的登陆域名来登录 VPN，

联通用户请登录：https://vpn.nscc-tj.cn

电信用户请登录：https://vpn1.nscc-tj.cn

联通电信用户，除了以上地址不同外，其它所有操作均相同，因此后续以联通接入用户为例，介绍其 VPN 登陆和数据传输方式，电信用户只需要将其中的 https://vpn.nscc-tj.cn 地址更换为 https://vpn1.nscc-tj.cn 即可。用户可以根据自己登陆 PC 的操作系统不同选择不同的系统登陆 VPN 方式，详述如下：

2.2.1 Windows 系统

用户采用浏览器方式登陆 VPN，不需要用户安装客户端软件。**推荐使用 IE 浏览器（或基于 IE 核心的浏览器，诸如遨游、360、搜狗等均可）。**

VPN 登陆步骤如下：

1. 在 IE 浏览器的“工具”——“Internet 选项”——“安全”——“可信站点”——“站点”选项，添加 https://vpn.nscc-tj.cn 站点（电信用户为 https://vpn1.nscc-tj.cn），如下图所示：

图 2-1 受信任站点

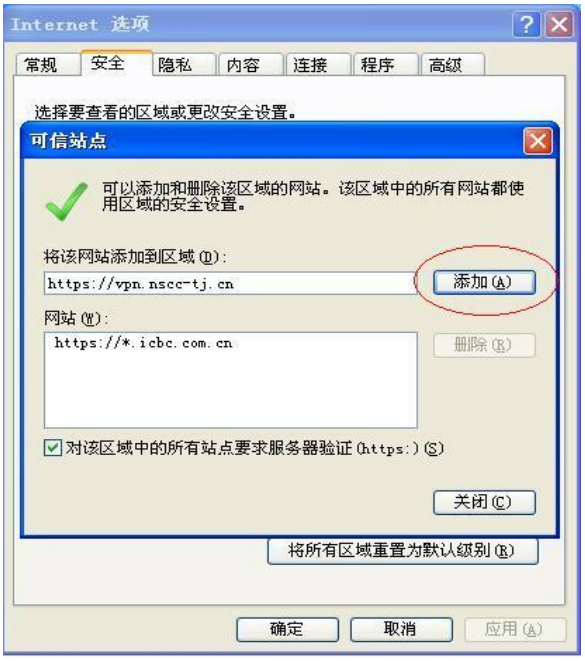

图 2-2 添加 VPN 站点

2. 点击“受信任站点”的自定义级别，确认 ActivX 选项是否启动。如下图所示：

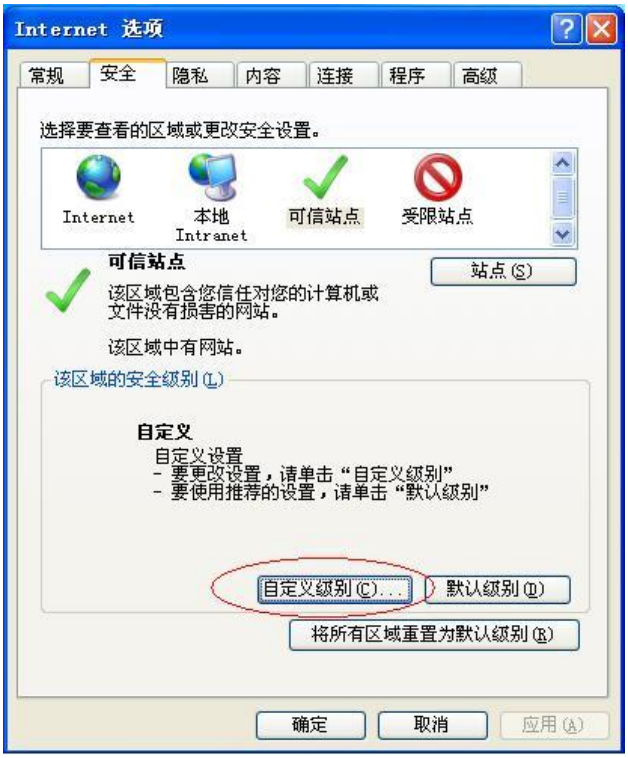

图 2-3 自定义级别

图 2-4 ActiveX 空间启用

3. 使用浏览器访问 https://vpn.nscc-tj.cn，可能会遇到如下图所示的提示，

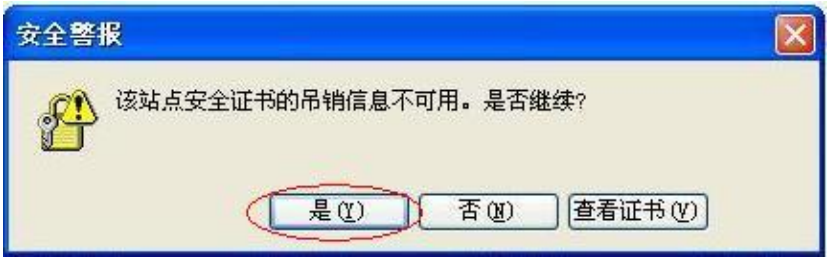

图 2-5 安全警告 1

图 2-6 安全警告 2

4. 均**点击红圈**所选项，从而进入如下界面

图 2-7 VPN 主页

5. 输入您的 VPN 用户名和密码，及验证码登陆至中心的 VPN，过程中还会遇

到如下提示：

图 2-8 控件安装提示

6. 请**点击安装相应的控件**，安装完成后重新登录后，进入如下界面

图 2-9 登陆后界面

如上图所示中，**TCP 应用为用户可以使用的资源，TH-1A 大系统的用户会具有 TH-1A-LN1，TH-1A-LN2，TH-1A-LN3 的 TCP 应用资源（如红圈所标识），分别表示 TH-1A 大系统的 LN1-LN3，三个登陆节点。**

以上过程即完成了 Windows 用户 VPN 的登陆。

2.2.2 Linux 系统

环境要求：

1) Linux 系统登陆 VPN 之前请确认系统中是否安装了 JRE 1.5 或更高版本，如果没有请自行安装。

2) Linux 系统登陆 VPN 必须是 root 用户，而且需要关闭 ssh 服务（可以通过命令：service sshd stop）。

登陆过程如下所示（以 Centos 5.5 为例）：

1．打开 Firefox 浏览器，打开“Edit > Preferences > Applications”，选择

JNLP file。在动作下拉菜单中选择“Use Other”（如图 2-10 所示）。

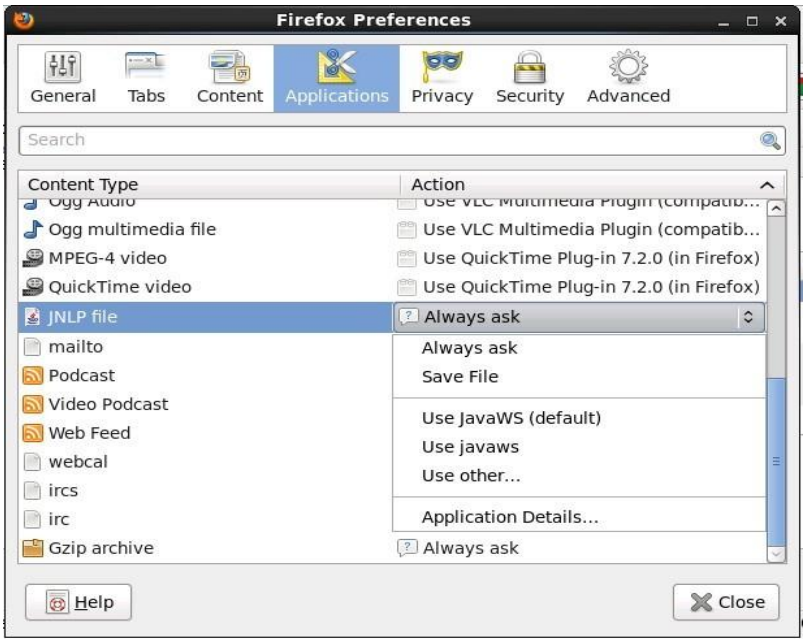

图 2-10 Firefox 配置 1

2．将位置设置为/usr/java/jre1.6.0_27/javaws/javaws，**其中/usr/java/jre1.6.0_27 为 JRE 的安装路径**，如图 2-11 所示。

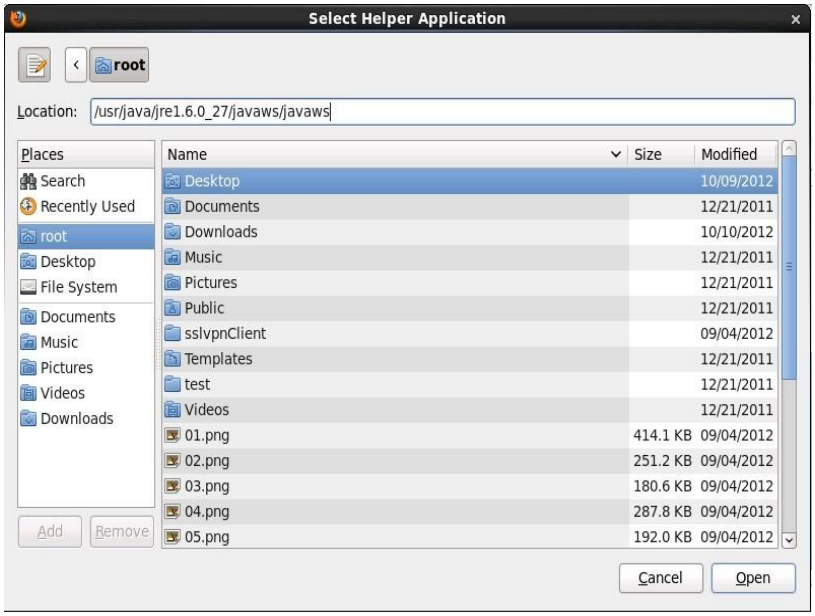

图 2-11 Firefox 配置 2

3．在 Firefox 浏览器的地址栏输入前面提到的 VPN 登陆地址（如联通 https://vpn.nscc-tj.cn），打开 SSL VPN 用户登录页面。输入用户名、密码以及

验证码后点击登陆。登陆界面如图 2-12 所示：

图 2-12 登陆界面

4. 如果出现“警告–安全”窗口，单击<Run>按钮。如果出现打开 cgi 文件的提示，选择<OK>，如图 2-13 所示：

图 2-13 cgi 文件提示界面

5. 如果出现“警告–安全”窗口，单击<Run>按钮。登陆成功后屏幕上方工具栏会出现一个 SSL VPN Client（如图 2-14 所示）。

图 2-14 登陆成功

注意：如果没有配置通知区域，则客户端不会显示上述图标，但不影响用户使用。

2.2.3 Mac 系统

目前为了满足 TH-1A 部分用户希望从 Mac 系统登陆 TH-1A 的要求，中心与设备供应商共同探讨出了目前的临时解决方案，用户可以通过 web 认证的方式，认证过后，保持该 web 页面即可以登陆 TH-1A 的登陆节点。具体的使用方法如下（以 Windows 下 IE 登陆为例，Mac 可以用自带的浏览器）：

1．打开浏览器，如图 2-15 所示，网通用户在地址栏输入 60.29.219.3，电信用户输入 123.150.4.85，回车。

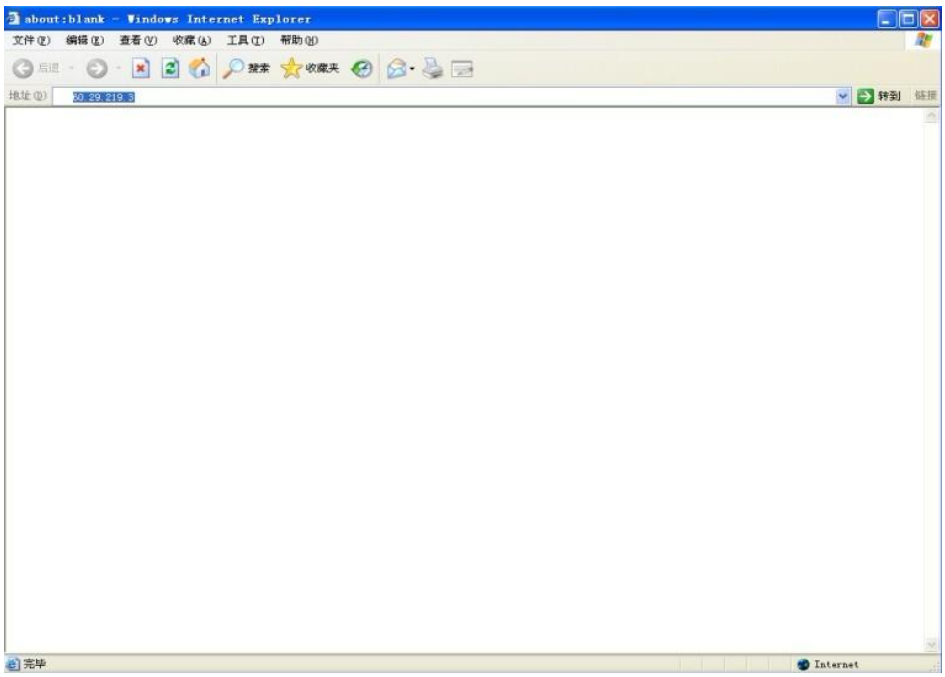

图 2-15 IMC 登陆地址

2．浏览器会出现一个登陆界面，如图 2-16 所示，按照给您的账户名和密码输入进去，服务类型不用管，点击上线。

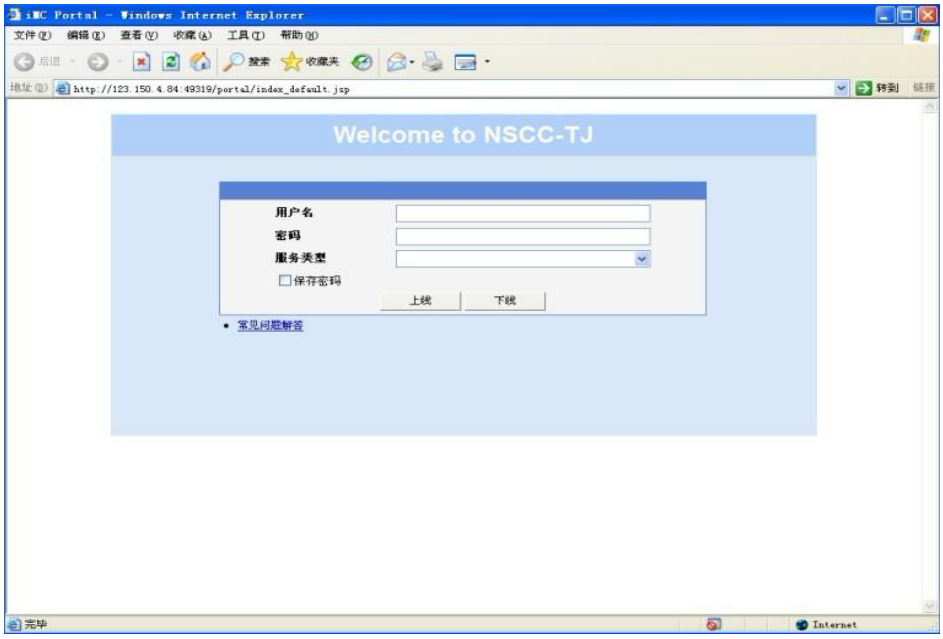

图 2-16 IMC 登陆界面

3．正确连接后会出现如下图 2-17 所示的界面，使用过程中不要关闭此界面。

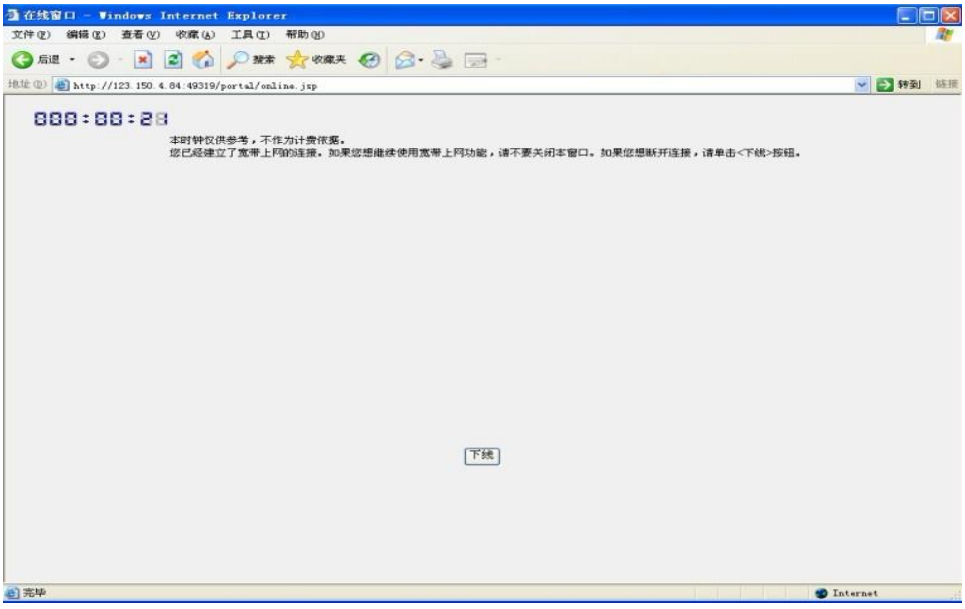

图 2-17 登陆成功

登陆成功后能 ping 通中心的登陆节点 IP，可以通过远程终端，**联通用户在 Host Name 一栏输入 60.29.219.3，电信用户在 Host Name 一栏输入 123.150.4.85，**

连接到中心的登陆节点。如果您需要断开本地 PC 与中心服务器的连接，先关闭远程终端，然后点击图 2-17 中的下线，即可正常退出。

注意：如果输入两个 IP 都无法跳转到登陆界面，可以直接输入如下 portal 地址：http://123.150.4.84:49319/portal/index_default.jsp

2.2.4 VPN 登陆注意事项

VPN 登陆注意事项如下：

1. 在国家超级计算天津中心网站 **http://www.nscc-tj.gov.cn** 右上角有“用户 VPN 登陆”选项，用户也可以从那里点击进入（联通与电信用户请分别选择适合自己的网络接入）。

2. 如果您在 Windows 下安装有 360 安全卫士等防护软件，**当有提示时，请点允许**，来保证 VPN 的正常使用。

3. VPN 用户默认限制，**同时使用上限为 5 人**。（一台 PC 机只允许登录一个 VPN 账号）如果您有特殊需求，需要更多的人同时使用，请告知中心。**由于 VPN 账号同时允许多人登陆，因此不允许修改密码，您如果需要修改密码，请联系中心技术人员。**

4. 在登陆 VPN 后，通过软件登陆系统，进行编译、提交任务等操作时，**请不要关闭浏览器或退出 VPN**，否则会断开连接。

5. 系统默认用户登陆 VPN 后，**如不进行任何操作（登录系统等操作），30 分钟后会自动下线。（如果您通过 ssh 软件连接天河登陆节点则不会自动下线）**

6. **如果您需要退出 VPN 时，请点击 VPN 右上角的“退出”，而不要直接关闭浏览器。（关闭浏览器系统会保留该用户 5 分钟，如果在该段时间内超过 VPN 用户限制 5，您将短时无法登陆）**

7. 您在使用中遇到问题可以参考“4.1 小节的 VPN 登陆问题解决”，如果还无法解决则可以 email 或电话咨询。**（相关联系方式见“5 小节的 技术支持”）**

2.3 登陆服务器和数据传输

2.3.1 登陆服务器

按照以上方式成功登陆中心的 VPN 后，用户则可以通过 ssh 服务登陆天河系统登陆节点来使用中心资源。为了保证用户的数据安全，中心不提供 telnet 等其他连接方式。

中心资源通过 **TCP 应用**的方式供用户使用，如图 2-9 所示，用户成功登陆 VPN 后，可以看到自己允许使用的资源。TH-1A 大系统具有 **TH-1A-LN0，TH-1A-LN1，TH-1A-LN2，TH-1A-LN3 的 TCP 应用，**用户可以使用 ssh 客户端软件（如 SSH Secure Shell Client，SecureCRT，Putty）来登录系统。SSH Secure Shell Client，SecureCRT，Putty 等均为免费软件，网络上均有下载。

登录时，Host Name 项填写 **TH-1A-LN2**，分别以 Secure Shell Client，SecureCRT，Putty 为例，登陆方式如下：

SSH Secure Shell Client 登陆界面如图 2-18 所示：

图 2-18 SSH Secure Shell Client 登陆界面

SecureCRT 登陆界面如图 2-19 所示：

图 2-19 SecureCRT 登陆界面

Putty 登陆界面如图 2-20 所示：

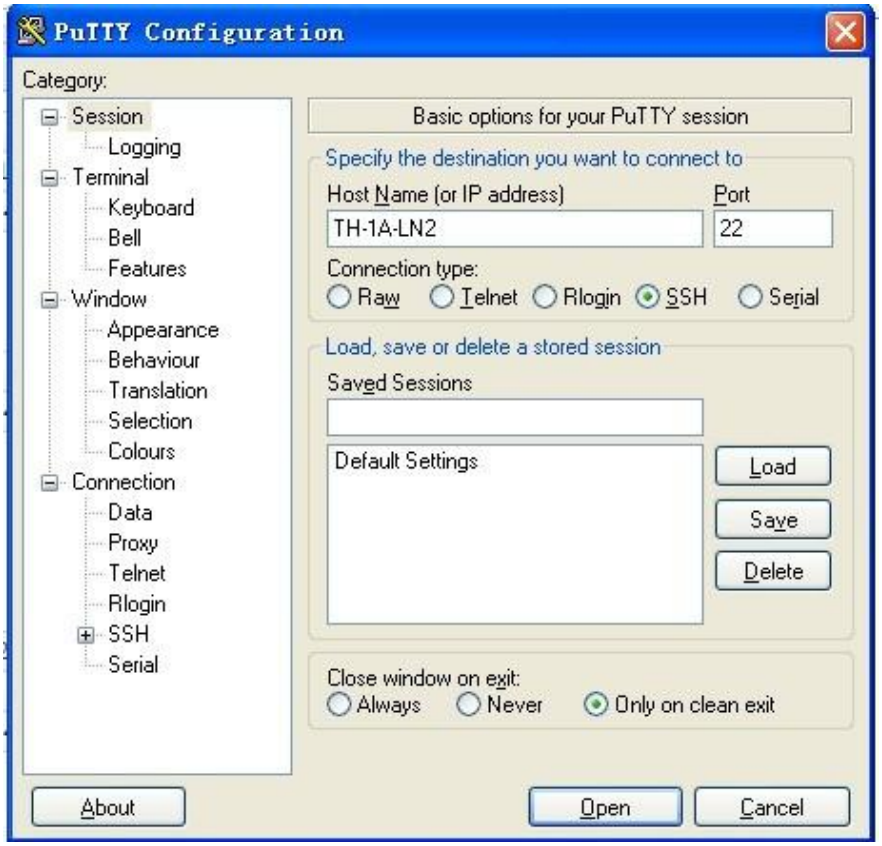

图 2-20 Putty 登陆界面

登陆之后，输入相应的用户名，密码，即可登陆至 TH-1A 大系统的 LN2 节点。

登陆 LN2 后，会收到如下提示：

```
1.Welcome to TH-1A System of NSCC-TJ.
2.If you have any problem, you can send mail to support@nscc-tj.gov.cn
```

之后您即可以开始编译、提交任务等操作。

特别注意：TH-1A 大系统的 LN0-LN3 为登陆节点，只负责用户的登陆，编译、提交任务等操作，不允许直接在 LN0-LN3 运行可执行程序。（详细描述见“1.1.1 小节”）

2.3.2 文件传输

目前 TH-1A 大系统只有 LN3 提供数据的上传、下载服务。**Linux 和 Mac 用户可以直接使用 scp 等命令拷贝数据，此处不再详述。**

Window 用户：从外部客户端向 TH-1A 大系统中上传或下载文件，可以使用 sftp 客户端，例如 SSH Secure Shell Client 等本身自带的文件传输功能，如图 2-21 所示：

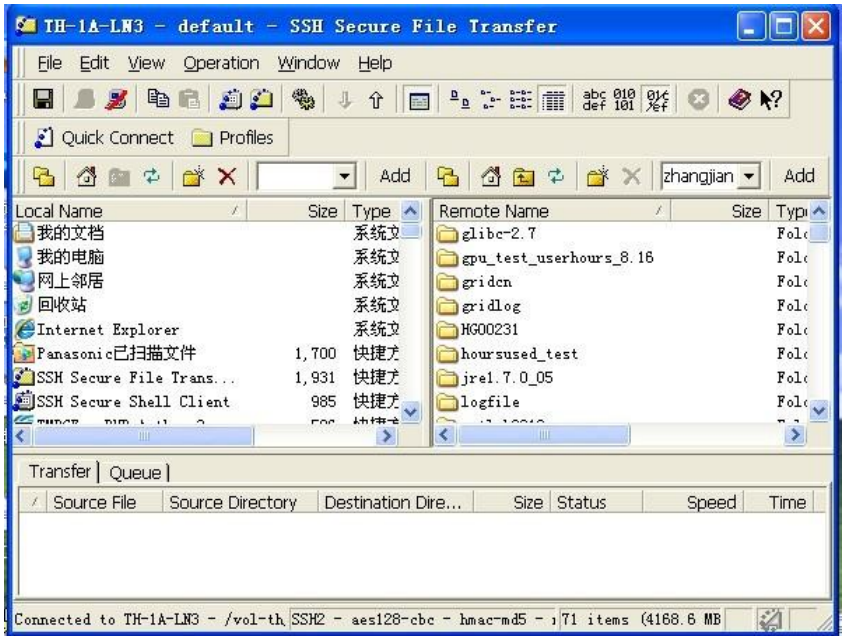

图 2-21 SSH Secure Shell Client 文件传输界面

或者使用 WinScp 的 sftp 数据传输软件（**免费软件，网络容易下载，且该软件支持断点续传，推荐使用**），登陆界面如图 2-22 所示：

图 2-22 WinSCP 登陆界面

以上软件即可实现文件传输功能。（**推荐使用 WinSCP 传输软件**）

2.4 环境变量设置

根据用户帐号使用的 Shell 的不同，设置环境变量的方法也有所不同。假设我们要增加一个用来表示字符串“/usr/local/bin”的环境变量 MYENV，可以采用下面的方法来设置。（**TH-1A 大系统默认用户选择的环境变量为 Bash**）

1）Bash 的设置方法

```
export MYENV=/usr/local/bin
```

如果需要环境变量在登录进用户帐号后自动设置，则可以编辑用户帐号起始目录（$HOME）下的.bashrc 文件，将上述命令行加入文件中。

2）sh 的设置方法

```
MYENV=/usr/local/bin
```

```
export MYENV
```

如果需要环境变量在登录进用户帐号后被自动设置，则编辑用户帐号起始目录（$HOME）下的.profile 文件，将上述命令行加入文件中。

3）csh 的设置方法

```
setenv MYENV /usr/local/bin
```

如果需要环境变量在登录进用户帐号后被自动设置，则编辑用户帐号起始目录（$HOME）下的.cshrc 文件，将上述命令行加入文件中。

2.5 退出系统

执行“exit”命令或按“ctrl-d”键，即可退出系统。

2.6 用户帐号密码修改

目前系统采用 LDAP 进行用户管理，新创建的用户第一次登陆节点时会创建相应的工作目录。用户可以通过 passwd 命令修改用户密码，以 ncps 用户为例，举例说明如下：

```
[ncps@ln2 ~]$ passwd
Changing password for user ncps.
Enter login(LDAP) password:
New password:
Re-enter new password:
LDAP password information changed for ncps
passwd: all authentication tokens updated successfully.
```

首先需要输入中心给分配的账户密码，之后再输入新的密码，重复输入一次后，就会显示密码更新成功。

特别提示：为了保证您用户的数据安全，您需要保证您的系统用户密码不外泄，希望您能经常更换系统用户密码（两个月更换一次为宜）。

如需更换 VPN 账号密码，请告知中心技术人员，我们帮您更换。

3 作业提交

TH-1A 大系统上的作业管理系统**以 CPU 核作为并行作业的资源分配单位**，实现并行作业的调度运行。在 TH-1A 大系统中，所有在计算节点中运行的串行或并行应用程序，都必须通过资源管理系统来提交运行。资源管理系统首先将用户提交的应用程序构造成作业进行排队处理，然后根据 TH-1A 大系统的实时运行资源状态，决定何时以及在哪些计算节点中加载应用程序的运行，不同的应用程序之间不存在资源的竞争冲突，用户也可以通过作业管理系统来监控应用程序的运行。

但为了保证系统资源的高效使用，用户请求的快速响应，系统的稳定性，在系统中做出了相应的使用限制，相关限制如下：

3.1 使用限制

3.1.1 分区限制

目前 TH-1A 大系统，根据用户的使用情况，将所有计算资源主要分成了三个区，如下表所示：

表 3-1 分区限制

分区名称	分区含义	分区限制		
		任务使用最多节点数	最多核数	任务最长运行时间（小时）
TH_SR	包机时用户分区	无	无	无
gpu_test	GPU 测试分区	64	768	2*24
TH_NET	普通用户区	343	4116	2*24
TH_NEW	长队列用户区	256	3072	10*24
debug	用户调试分区	2	24	0.5

不同分区针对不同的用户群体开放使用，用户可以使用 yhi -l 命令，看到相应的分区限制信息。

其中 PARTITION 表示分区，TIMELIMIT 表示该分区的时间限制，NODES 表示节点数，STATE 表示节点运行状态其中 down 表示未启动，idle 表示启动后处于空闲状态，allocated 表示节点已经分配了一个或多个作业，NODELIST 为节点列

表。

所有分区均可以设定相应允许的用户队列，目前 TH_NET 分区为创建账号默认队列，最多提供 2*24 小时的计算服务；TH_NEW 分区为长队列，最多提供 10*24 小时的计算服务；gpu_test 分区供用户测试 GPU 程序等使用，单个任务最多使用 64 个节点，共计 768 核。中心根据用户的不同分类，划分不同的资源，您如果看不到某些分区，是因为您不具备相应的资源使用权限。

注意：

1. 由于大型集群系统具备一定故障率，TH-1A 大系统系统十分庞大，为了保证系统稳定性，分区中有限定任务执行时间的限制，**因此建议用户为程序设立“断点”**从而保证任务由于意外中断后，可以继续运算。
2. 如果您的程序没有办法“续算”，而且运行时间超过 2*24 小时，请联系中心技术人员。
3. TH_SR 主要针对包机时用户提供服务，请包机时用户后续使用该分区的资源。该分区还提供用户预约，如果包机时用户计划跑大规模作业，请用户提前 2-3 天联系我们，我们会提前做好准备。
4. debug 是用户调试分区，每个用户都可以使用最大 2 个节点 24 核的资源，作业时间限制为 30 分钟。

3.1.2 用户限制

除了上述的分区限制，目前还根据用户的申请情况，针对用户做了一定的限制，**该限制主要基于用户和中心签订合同的规模**。

包括：最多可以使用的节点数、最多可以使用的核数、单个任务最多可以使用的节点数、单个任务最多可以使用的核数等。

用户在使用过程中，如果有超出自己合同范围内的计算规模的计算需求，请基于自己的需求，向中心提出申请，中心会根据用户需要审查后，进行一定的修改。

为了保证系统和用户数据的安全，目前普通用户不能在没有申请资源时，就

ssh 链接到计算节点，只有分配了相应的计算节点资源后，才能 ssh 到指定计算节点。

3.1.3 磁盘配额限制

为了合理利用有限的存储资源，目前中心对用户默认进行存储软限制 500G，存储硬限制 1T，文件数软限制 100 万，文件数硬限制 200 万的磁盘配额限制。用户登录后会出现如图 3-1 的磁盘配额信息：

```
Disk quotas for group zhangjian (gid 1101):
     Filesystem  kbytes   quota   limit   grace   files   quota   limit   grace
       /vol-th 16540636  524288000 1073741824       -  234355  1000000 2000000       -
```

图 3-1

图 3-1 的参数从左往右看，“Filesystem”对应用户所在的共享分布式存储，kbytes 对应的是用户目前已经使用的存储（单位 KB），“quota”对应磁盘软限制（单位 KB），“limit”对应磁盘硬限制（单位 KB），“grace”对应磁盘存储状态，“files”对应用户现有的文件数，“quota”对应文件数软限制，“limit”对应文件数硬限制，“grace”对应文件数状态。

以磁盘存储为例说明软、硬限制的含义，文件数软、硬限制的含义与其一样。用户使用存储低于 500G 时，如图 3-1 所示，存储状态正常；当用户使用存储介于 500G 和 1T 之间时，存储状态如图 3-2 所示，kbytes 参数对应的数字带有“*”表示用户配额异常，“4w1d23h59m57s”表示一个月的倒计时，如果用户在倒计时结束前将使用存储清理到 500G 以下，则存储状态恢复正常，否则如图 3-3 所示，用户存储无法写入，如图 3-4 所示；如果用户使用存储大于 1T，配额信息如图 3-3 所示，操作时会受限制，如图 3-4 所示。

```
Filesystem   kbytes    quota    limit    grace    files
   /vol-th 16548056* 1048576 1073741824 4w1d23h59m54s
```

图 3-2

```
Filesystem   kbytes    quota    limit    grace
   /vol-th 16548056* 1048576 1048576        -
```

图 3-3

```
[zhangjian@ln2%tianhe ~]$ cp glibc-2.7.tar.gz quota_test/
cp: writing `quota_test/glibc-2.7.tar.gz': Disk quota exceeded
cp: closing `quota_test/glibc-2.7.tar.gz': Input/output error
```

图 3-4

注意：有的时候用户登录会出现错误提示“Some errors happened when getting quota info. Some devices may be not working or deactivated. The data in "[]" is inaccurate.”这是因为登陆节点 quota 服务没有启用，对用户本身的操作和作业不会有影响。

用户可以用命令“lfs quota -g username /vol-th”随时查看自己的配额信息。

3.2 状态查看命令

在用户提交作业前，应查看系统的使用情况，这样利于用户根据系统使用情况，进行选择，例如只是做调试编译，则可以使用 debug 分区，这时候可以通过状态查看命令获取信息，选择相应计算节点。

3.2.1 节点状态查看 yhinfo 或 yhi

yhi 为 yhinfo 命令的简写，用户可以使用 yhi 或者 yhinfo 命令查看节点的使用情况，从而根据情况做出选择。

其中 PARTITION 表示分区，TIMELIMIT 表示该分区的时间限制，NODES 表示节点数，STATE 表示节点运行状态其中 down 表示未启动，idle 表示启动后出于空闲状态，allocated 表示节点已经分配了一个或多个作业，NODELIST 为节点列表。

3.2.2 作业状态信息查看 yhqueue

yhqueue 或 yhq 命令用于查看系统中，各计算节点的运行情况，如图 3-5 所示：

```
[fengjh@ln2%th1 ~]$ yhq
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
   1639    normal   111-M1   wanggc   R    4:11:08      1 cn191
   1638    normal    o00C5     tyan   R    4:25:14      1 cn142
    742    normal  wld-v11     simg   R 1-04:53:10      1 cn164
   1635    normal    b00C5     tyan   R    4:52:32      1 cn141
   1640    normal VAH=VA+H   wanggc   R    4:11:08      1 cn191
    787    normal    d00C3     tyan   R   21:49:38      1 cn184
    788    normal    d00C3     tyan   R   21:49:38      1 cn58
   1637    normal    o00C5     tyan   R    4:50:50      1 cn57
   1634    normal    b00C5     tyan   R    5:04:10      1 cn55
   1633    normal c2h5oc2h   wanggc   R    5:05:12      1 cn187
   1632    normal   TSR2-2   wanggc   R    5:09:28      1 cn187
   1629    normal    d00C5     tyan   R    5:27:06      1 cn138
   1628    normal    d00C5     tyan   R    5:27:21      1 cn137
   1626    normal   TSR2-1   wanggc   R    5:34:21      1 cn177
   1625    normal   TSR1-4   wanggc   R    5:37:07      1 cn177
   1624    normal   TSR1-3   wanggc   R    5:39:12      1 cn165
   1623    normal   TSR1-2   wanggc   R    5:43:27      1 cn165
   1615    normal ethane-2   wanggc   R    6:23:49      1 cn54
```

图 3-5 节点运行任务状态信息

JOBID 表示任务 ID，Name 表示任务名称，USER 为用户，TIME 为已运行时间，NODES 表示占用节点数，NODELIST 为任务运行的节点列表。**获取的 jobid，用户在作业取消命令 yhcancel 中会使用到。**

用户可以使用 yhq 查看自己提交的作业，为了保证用户的数据安全，普通用户通过 yhq 只能看到自己提交的作业。

查看作业明细：

用户可以通过如下命令来查看自己提交的作业明细

```
yhcontrol show jobs jobid
```

其中 jobid 表示作业的 id 号，用户根据自己作业的情况填入即可，之后用户即可以看到该作业十分详细的信息。

3.3 提交作业

目前 TH-1A 大系统部署的资源管理系统包括多种作业提交方式，包括批处理作业提交方式 yhbatch，交互作业提交方式 yhrun 和分配模式 yhalloc。**作业终止方式为 yhcancel 命令，需要获取作业的 jobid，如前所述，jobid 可以通过 yhq 命令查看获得。**

本手册，为了简化和方便用户，只对相关命令做简单介绍，用户如需更多参数选择，则可以通过响应命令后加入—help 的方式，获取帮助信息，从而满足用户需求。

3.3.1 批处理作业 yhbatch

注意：如果没有交互需求，请使用 yhbacth 提交任务。yhbatch 提交的作业终端关闭时不会受到影响，登陆节点 down 机时也不会受到影响，强烈推荐使用 yhbatch 提交任务。

yhbatch向资源管理系统提交一个批处理脚本，yhbatch将在脚本成功提交到资源管理系统控制进程并分配作业JobID 后立即退出。

批处理脚本可能不会被立刻分配资源，而是在排队作业队列中等待，直到资源需求得到满足。当批处理脚本被分配资源后，资源管理系统将在所分配的第一个节点上运行批处理脚本。

yhbatch 运行的主要格式如下：

```
yhbatch [options] program
```

yhbatch 包括多个选项，用户最常使用的选项如下：

-n, --ntasks=ntasks 指定要运行的进程数。请求 yhrun 分配/加载 ntasks 个进程。省缺的情况是每个 CPU 运行一个进程，但是-c 参数将改变此省缺值。
-N, --nodes=minnodes[-maxnodes] 请求为此作业至少分配 minnodes 个节点。调度器可能决定在多于 minnodes 个节点上启动作业。可以通过指定 maxnodes 限制最多分配的节点数（如“--nodes=2-4”）。最少和最多节点数可以相同以便指定确切的节点数（如“--nodes=2-2”将请求两个并且仅仅两个节点）。如果没有指定-N，省缺的行为是分配足够的节点以满足-n 选项的要求。
-p, --partition=partition 从分区 partition 请求资源。如未指定，则省缺为默认分区。
-t, --time=minutes 设置作业的运行时间限制为 minutes 分钟。省缺值为分区的时间限制值。当到达时间限制时，作业的进程将被发送 SIGTERM 以及 SIGKILL 信号终止执行。
-D, --chdir=path 加载的作业进程在执行前将工作目录改变到 path。省缺情况下作业 yhrun 进程的当前工作目录。
-l, --labe 在标准输出/标准错误的每行之前添加任务号。通常，远程任务的标准输出和标准错误通过行缓冲直接传递到 yhrun 的标准输出和标准错误。--label 选项将在每行输出前面添加远程任务的 ID。
-J, --job-name=jobname

指定作业的名字。省缺值是可执行程序的名字 program。
-W, --wait=seconds 指定在第一个任务退出后，到终止所有剩余任务之前的等待时间。0 表示无限等待（60 秒后将发出一个警告）。省缺值可由系统配置文件中的参数设置。此选项用于确保作业在一个或多个任务提前退出时能够及时终止。
-w, --nodelist=nodelist|filename 请求指定列表中的节点。分配给作业的将至少包含这些节点。nodelist 可以是逗号分割的节点列表或范围表达式（如 cn[1-5,7,12]）。如果包含“/”字符，则 nodelist 将会被当作是一个文件名，其中包含了所请求的节点列表。
-x, --execlude=nodelist|filename 排除指定列表中的节点。分配给作业的将不会包含这些节点。
--checkpoint-path=path 指定任务检查点映像文件的保存目录。省缺为任务的当前工作目录。
--checkpoint-period=number[h|m] 指定对作业进行自动周期性检查点操作。如果 number 后没有跟时间单位，则默认为 h（小时）。
--restart-path=path 指定本次任务加载为从以前的检查点映像恢复执行。path 为检查点映像文件所在的路径。
--exclusive 此作业不能与其它运行的作业共享节点，加入此选项，则表示用户需要针对此作业使用独占的处理器，如果没有足够的处理器，则作业的启动将会被推迟。

以上选项中，由以－N，-n，-p，-w，-x 等选项最常用，-N 指定节点数，-n 指定进程数，-p 指定分区名，-w 指定节点列表，-x 指定不参加分配的节点列表（用于排除自己认为有问题的节点）

用户在 yhbatch 的参数中指定资源分配的需求约束，编写的作业脚本中，**也可以使用 yhrun 命令加载计算作业，此时 yhrun 通过环境变量感知已经分配了资源，从而直接创建作业而不再次提交作业**。

批处理作业的脚本为一个文本文件，脚本第一行以“#！”字符开头，并制定脚本文件的解释程序，如 sh，bash，rsh，csh 等。

这种作业提交方式，适合提交绝大多数作业。如果需要连续执行多个任务的作业，用户可以在脚本中提交多个任务，逐个计算。

如前所述，系统中作业的运行分成**两步：资源分配与任务加载**。批处理作业

使用 yhbatch 提交脚本的方式运行，yhbatch 负责资源分配，yhbatch 获取资源后，会在获取资源的第一个节点运行提交的脚本。

举例一如下：

假设用户作业为可执行文件 run.sh，编写提交脚本 sub.sh 如下：

```
#!/bin/bash
yhrun -n 16 -p TH_NET run.sh
```

然后给 sub.sh 脚本增加可执行权限：

```
chmod +x sub.sh
```

最后根据用户的资源需求，提交如下：

```
yhbatch -n 16 -p TH_NET ./sub.sh
```

计算完成后，工作目录中会生成以 slurm 开头的.out 文件为输出文件。

注意：**yhbatch 申请的资源必须不小于 sub.sh 脚本中 yhrun 申请的资源。**

举例二

yhbatch 提交的脚本中即可以包含 yhrun，也可以支持 mpirun 等提交作业方式。例如使用了/vol5/mpi-gcc/openmpi-1.4.3 编译生成可执行程序 a.out，需要运行在节点 cn12-cn27，共计 16 个节点 128 个进程。则安装 mpirun 提交任务的规则，需要撰写 **hostlist 文件**包含 cn12-cn27，如下所示：

```
cn12:8
cn13:8
cn14:8
cn15:8
cn16:8
cn17:8
cn18:8
cn19:8
cn20:8
cn21:8
cn22:8
cn23:8
```

```
cn24:8
cn25:8
cn26:8
cn27:8
```

之后撰写脚本 sub.sh 如下：

```
#!/bin/bash
/vol5/mpi-gcc/openmpi-1.4.3/bin/mpirun －hostfile hostlist －np 128 ./a.out
```

用户根据该脚本（chmod 修改该脚本可执行权限 chmod +x sub.sh），提交批处理命令如下：

```
yhbatch －N 16 －p test －w cn[12-27] ./sub.sh
```

注意：TH-1A 系统上的资源使用抢占式调度方式，即作业在节点上哪怕只运行了一个核的进程，其他作业也无法再分配到该节点上。

特别提示：批处理作业提交模式，试用范围很广，由于手册篇幅限制，不能详述，如果您在提交批处理作业的过程中遇到了任何问题，请联系中心技术人员。

3.3.2 交互式作业提交 yhrun

对于**交互式作业**，资源分配与任务加载两步均通过 yhrun 命令进行：当在登录 shell 中执行 yhrun 命令时，yhrun 首先向系统提交作业请求并等待资源分配，然后在所分配的节点上加载作业任务。

yhrun 运行的主要格式如下：

```
yhrun [options] program
```

yhrun 包括多个选项，与 yhbatch 类似。

示例：

1）**在分区 TH_NEW，节点 cn[4-16] 上运行 hostname**

```
$ yhrun -w cn[4-16] －p TH_NEW　hostname
yhrun: XXXXX: use ‘-t’ option to set time limit of job. defaults to 5 (minutes)
yhrun: job 4385 queued and waiting for resources
```

```
yhrun: job 4385 has been allocated resources
cn4
cn7
...
cn14
```

2）运行在 gpu_test 分区，运行 4 任务的 MPI 程序 cg.C.4，每个节点一个任务，分配的节点中至少包含节点 cn[4-5] ；作业运行时间不超过 20 分钟；运行过程中查看任务状态

```
$ yhrun -w cn[1-2] -n 4 -N 4 -t 20 －p debug cg.C.4
NAS Parallel Benchmarks 3.2 --CG Benchmark
Size: 150000
Iterations: 75
Number of active processes: 4
Number of nonzeroes per row: 15
Eigenvalue shift: .110E+03
iteration ||r|| zeta
1 0.15244429457374E-12 109.9994423237398
2 0.45529118072694E-15 27.3920437146522
3 0.45039339889198E-15 28.0339761840269
4 0.44936453849220E-15 28.4191507551292
yhrun: interrupt (one more within 1 sec to abort)
yhrun: task[0-4]: running
5 0.44884028024712E-15 28.6471670038895
6 0.44551302644602E-15 28.7812969418413
```

特别注意：

1. yhrun 基本可以替代 mpirun，特别是使用/usr/local/mpi 或/usr/local/mpi-gcc 目录下 mpi 编译的程序，完全可以使用 yhrun 提交任务，而不需使用 mpirun。

2. yhrun 为交互式作业提交方式，用户如需要和程序进行交互，则选择直接使用 yhrun 提交任务，**如果不需要交互，则需使用批处理作业提交方式。**

3. yhrun 提交的任务，如果没有进行输入输出的重定向，在关闭登陆客户端软件时，会导致任务中断，因此如无特殊需要，请直接使用 yhrun 提交任务时，重定向输入输出，并保留相应的 log 文件，方便遇到问题时，技术人员及时解决。

重定向举例如下：

```
yhrun －p test －N 16 －n 128 ./a.out >log 2>&1 &
```

>为重定向符号，2>&1 表示标准错误输出重定向至标准输出，最后的&表示后台提交方式，这样保证了该任务在登陆客户端关闭时依然保持不中断。

4. 再次提示，为了保证任务的稳定性，如无特殊需要请使用批处理作业 yhbatch 提交方式。

3.3.3 分配模式作业 yhalloc

分配作业模式类似于，交互式作业模式和批处理作业模式的融合。用户需要指定资源分配的需求条件，向资源管理器提出作业的资源分配请求。作业排队，当用户请求资源被满足时，将在用户提交作业的节点上，执行用户所指定的命令，指定的命令执行结束后，也运行结束，用户申请的资源被释放。

yhalloc 后面如果没有跟定相应的脚本或可执行文件，则默认选择了/bin/sh，用户获得了一个合适环境变量的 shell 环境。

yhalloc 和 yhbatch 最主要的区别是，yhalloc 命令资源请求被满足时，直接在提交作业的节点执行相应任务。而 yhbatch 则当资源请求被满足时，在分配的第一个节点上执行相应任务。

yhalloc 在分配资源后，再执行相应的任务，很适合需要指定运行节点，和其它资源限制，并有特定命令的作业。例如 ansys 或其他工程仿真软件的模块，以 ansys 的 lsdyna 模块为例，在并行计算机系统中，lsdyna12.1 版本，需要指定相应的 memory，相应的执行节点列表。由于用户需要在命令中指定相应计算节点，则适合用 yhalloc

例如：ansys 用户需要 8 个节点，32 个进程，每个节点 4 核的计算资源，利用 yhalloc，有两种提交方式。

第一种首先申请资源，执行如下命令：

```
yhalloc －N 8 －n 32
```

通过 yhq 查看相应的 jobID 为 163，节点为 cn[60-67],则用户可以选择如下方

式：

ssh cn60　　切换到 cn60 节点，之后执行如下命令：

```
lsdyna121 pr=dyna -dis memory=250m i=test.k o=test.out -machines cn60:4:cn61:4:62:4:63:4:64:4:65:4:66:4:67:4
```

则可以正常执行 lsdyna 程序。

第二种作业提交方式：

首先通过 yhi 命令，查看哪些节点空闲，确定 8 个空闲的节点，如确定的 8 个空闲节点为 cn[64-71]，则写如下脚本 lsdyna.sh：

```
#!/bin/bash
lsdyna121 pr=dyna -dis memory=250m i=test.k o=test.out -machines cn64:4:cn65:4:66:4:67:4:68:4:69:4:70:4:71:4
```

然后执行如下命令：

```
yhalloc －N 8 －n 32 －w cn[64-71] ./lsdyna.sh
```

使用如上方式，请注意，通过 chmod +x lsdyna.sh 给脚本加可执行权限。

yhalloc 包含多个选项，基本和 yhrun 类似，此处就不再详述，用户可以通过 yhalloc --help 命令查看相应所需参数。

特别提示：

1. yhalloc 和 yhbatch 的使用方法类似，主要区别为任务加载点不同，**yhalloc 命令资源请求被满足时，直接在提交作业的节点执行相应任务。而 yhbatch 则当资源请求被满足时，在分配的第一个节点上执行相应任务。**

2. yhalloc 提交的作业，如果需要关闭客户端，请重定向输入输出，并后台提交。

3.4 任务取消 yhcancel

yhcancel 取消任务，取消掉用户运行的任务，使用方式：

用户启动一个新的 ssh 连接，并执行 yhcancel JOBID 如图 3-3 所示：

```
[fengjh@ln2 ~]$ yhq
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
 118230       all    a.out   fengjh   R       0:06      8 cn[0-7]
[fengjh@ln2 ~]$ yhcancel 118230
[fengjh@ln2 ~]$ yhq
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
```

图 3-3 yhcancel 使用举例

yhcancel 命令强制取消任务后，显式的信息如图 3-4 所示：

```
yhrun: Force Terminated job 118230
slurmd[cn0]: *** STEP 118230.0 CANCELLED AT 2010-07-20T15:42:11 ***
yhrun: error: cn6: tasks 48-55: Terminated
yhrun: error: cn4: tasks 32-39: Terminated
yhrun: error: cn7: tasks 56-63: Terminated
yhrun: error: cn3: tasks 24-31: Terminated
yhrun: error: cn1: tasks 8-15: Terminated
yhrun: error: cn0: tasks 0-7: Terminated
yhrun: error: cn5: tasks 40-47: Terminated
yhrun: error: cn2: tasks 16-23: Terminated
yhrun: XXX: job done
```

图 3-4 任务取消后显示信息

4 常见问题

4.1 VPN 登陆问题

Q：Windows 系统登陆 VPN 无法加载插件或报错

A：按照用户手册确认是否打开了浏览器的允许加载插件选项，并确保您机器的安全软件、杀毒软件等不会阻止插件的正常运行。

Q：Windows 系统登陆 VPN 成功，但仍连接不上服务器

A：请确认您的桌面任务栏的通知区域里是否有“SSL VPN Client”图标。如果没有，则说明您没有登录成功，插件没有正常运行，请您参考手册中登陆部分重新进行设置确认；如果有，则需要进一步去确认您系统的主机名解析文件 C:\WINDOWS\system32\drivers\etc\hosts 内是否有 VPN 页面上的机器标识，如果没有，则说明您的该文件被保护或您当前的用户对该文件没有写权限，请您用系统管理员用户登录，以及关闭安全软件及杀毒软件对该文件的控制。

Q：登录 VPN 提示“超出人数限制”

A：VPN 用户默认最多允许 5 个人同时登陆，超过这个限制就会报错。但如果用户登录 VPN 后没有点击“退出”按钮正常退出，而是直接关掉页面，则会保持该次连接会话直到超时，此时也会占用一个登录连接数。

Q：Linux 系统登陆 VPN 不成功

A：linux 用户登录 VPN 需要使用 root 账户，并且同时关闭自己机器的 ssh 服务，常用版本的关闭方法为“/etc/init.d/sshd stop”。具体请参考用户手册或者 VPN 登陆下面的“SSL VPN 设置向导”。linux 系统的登陆与 windows 大致相同，但其需要 java 的支持，其对应的主机名解析文件为/etc/hosts。

Q：Mac 系统如何在 1 个 IP 下实现多台 Mac 系统接入“TH 系统”？

A：目前 MAC 系统登录的 VPN 认证是基于 IP 建立的，也就是说同一个 IP 地址只能建立起一个连接，一旦连接建立，使用同一个 IP 访问的客户端就都可以连接到系统上来了，其他客户端不需要再登录这个 VPN 进行认证了，但是一旦登录

VPN 的客户端退出，那么整个网络也就会断开了。

4.2 系统登陆问题

Q：登陆节点 home 目录下看不到原有的用户文件

A：这是由于登录节点启动后还没有挂载相应的共享存储，请用户先退出系统，稍等待后再重新进行登录。

Q：在登陆节点执行“ls；cd”这样的基本命令会卡死

A：这有可能是由于该登陆节点负载过大造成的，此时用户可以尝试更换其他登陆节点。

Q：登陆节点无法连通

A：这有可能是用户在登陆节点上运行非法程序导致节点宕机，我们会实时对系统进行监控，出现这种情况请用户更换其他登陆节点。建议用户不要在登陆节点上运行任何计算，一旦查到并影响到其他人的使用，则会进行警告，屡次不改者可能会被封号。

4.3 作业运行问题

Q：作业在某一个时间点后无输出

A：导致这种现象的原因很多，需要具体问题具体分析，其中一种可能是由于用户磁盘配额已满，无法写入数据造成的，因此需要用户及时清理自己的数据，如果并非这种情况请您邮件或者电话与我们的系统管理人员取得联系，我们将进行进一步查看。

Q：作业断开，slurm 日志中出现“DUE TO TIME LIMIT”报错信息

A：这是因为作业运行时间超过队列最大运行时间限制，请注意您所在队列的运行时间限制以及您作业已运行时间。

Q：作业断开，slurm 日志中出现“Not enough endpoint resources”报错信息

A：这是由于上一个作业结束时出现异常，节点 endpoint 没有正常释放。用户提交可以加-x 剔除问题节点，然后联系管理员进行解决。

Q：作业断开，slurm 日志中出现“Group ID not found on host”报错信息

A：这是由于计算节点的 passwd 和 group 没有与管理节点同步导致。用户提交可以加-x 剔除问题节点，然后联系管理员进行解决。

Q：作业断开，slurm 日志中出现“No such file or directory: going to /tmp instead”报错信息

A：这是由于计算节点没有挂载共享存储。用户提交可以加-x 剔除问题节点，然后联系管理员进行解决。

Q：作业断开，slurm 日志中出现“Job credential expired”报错信息

A：这是由于计算节点时间没有与管理节点同步。用户提交可以加-x 剔除问题节点，然后联系管理员进行解决。

Q：作业断开，slurm 日志中出现“bus error”报错信息

A：导致“bus error”的报错原因很多，具体问题需要使用工具排查。用户提交可以加-x 剔除问题节点，然后联系管理员进行解决。

Q：提交的作业总是被自动退出

A：您如果是使用 yhrun 提交任务，那么终端关闭、脚本终止都会导致任务被杀掉。建议用户使用 yhbatch 的提交方式，yhbatch 提交的任务，终端关闭甚至登陆节点宕机都不会对已提交的作业有影响。另外，还有可能是您提交的作业所分配的计算节点有问题导致自动退出，请您仔细查看产生的日志文件的报错信息，是否属于以上问题中的一种，并采取相应的处理。

Q：作业状态“S；CG”分别表示什么含义

A：“S”表示管理员将用户作业挂起以进行故障检测或故障处理，处理完后会将该作业恢复，不会对作业产生任何影响；“CG”是为作业结束或取消后的“退出”标记，不会对用户作业造成影响，请正常使用，如作业长时间处理“CG”状态，管理员会对其进行恢复处理。

Q：作业断开，slurm 日志提示找不到某个动态链接库

A：需要用户将动态链接库的路径添加到自己运行的环境变量中，假设缺少 x 库，先“locate x”找到该链接库的地址$DIR，然后编辑用户目录下的配置文件.bashrc，添加“export LD_LIBRARY_PATH=$DIR:$LD_LIBRARY_PATH”。

Q：查看有可用节点，但作业却一直处于 PD 状态

A： TH 系统的资源管理器采用“先进先出”的作业调度方式，作业处于 PD 状态说明在用户前面有其他用户先提交了作业，并且之前的用户作业超出了目前的可用资源总数，请用户耐心等待。根据用户资源需求，系统管理人员也会定期进行资源调整，降低作业排队时间。

4.4 存储问题

Q：登陆系统时提示“Some errors happened when getting quota info”

A： 这是由于在对系统进行调整时登陆节点 quota 服务没有启用导致，对用户本身的操作和作业不会有影响，管理员会定时对此进行调整，请您放心使用。

Q：默认的磁盘配额是多少？磁盘配额的含义是什么？

A： 为了合理利用有限的存储资源，目前中心对用户默认进行存储软限制 500G，存储硬限制 1T，文件数软限制 100 万，文件数硬限制 200 万的磁盘配额限制。以磁盘存储为例说明软、硬限制的含义，文件数软、硬限制的含义与其一样。用户使用存储低于 500G 时，存储状态正常；当用户使用存储介于 500G 和 1T 之间时，用户配额异常，通过“lfs quota –g username /vol-th”查看账号配额会看到已使用存储的数字旁边有一个“*”号，状态“4w1d23h59m57s”表示一个月的倒计时，如果用户在倒计时结束前将使用存储清理到 500G 以下，则存储状态恢复正常，否则，用户存储无法写入；如果用户使用存储大于 1T，用户会无法写入。

Q：磁盘无法写入，报“quota error”错误

A： 这是由于用户使用存储或文件数超过配额设定，需要用户对数据进行清理到磁盘配额软限制以下方可继续使用。

5 技术支持

由于用户手册篇幅有限，只列出了用户使用系统的基本方法以及常见问题和解决方法，很难面面俱到，还请您能够谅解。如果您在系统使用过程中遇到任何问题，都可以及时与中心技术人员取得联系。中心技术人员会在收到用户问题反馈后的 24 小时工作时间内给予回复。

1. 合同、资源申请使用、应用软件相关问题联系方式：

   Email：service@nscc-tj.gov.cn

   电话：022-65375561

2. 系统使用、作业运行相关问题联系方式：

   Email：support@nscc-tj.gov.cn

   电话：022-65375560，18302248223

重点提示：为了能尽快使您的问题得到定位解决，请您在通过邮件或电话联系中心技术人员时提供以下基本信息（建议拷贝表格填写）：

系统用户名	作业 jobid	作业日志路径	作业提交命令	错误现象

包括：系统用户名，出错作业号（jobid），出错作业日志路径，作业提交命令，错误现象描述等信息。

在此，特别感谢您对我们工作的信任与支持，同时也祝您在天河系统使用过程中工作愉快！希望我们能够共同携手推动我国并行计算技术的发展，合理使用并行计算资源在各个科研及工业领域不断创新突破！

附录 A 常用 Unix 命令

A1 基本命令

date：显示日期和当前时间，命令格式：$date。
who：查询当前登录在系统中的用户信息，命令格式：$who。
w：查询当前登录在系统中的用户行为，命令格式：$w。
write：将消息直接发送到另一个用户的终端上，命令格式：
$write username
Hello: We have a meeting at Room 412.
键入 Ctrl-D 结束输入消息，在 username 用户终端上可以看到上述信息。
mesg：选择是拒绝还是接受由 write 发来的消息，命令格式：
$mesg n 拒绝由 write 发来的消息；
$mesg y 允许别的用户发送消息；
$mesg 报告当前是否允许别的用户向你的终端发送消息。
ps：用于查看当前系统中的活跃进程，命令格式：$ps [options]。
kill：终止指定进程，命令格式：$kill [-signal] pid。

A2 目录操作

mkdir：创建目录，命令格式：$mkdir directory ...。
rmdir：删除目录，命令格式：$rmdir directory ...。。。
pwd：显示当前工作目录，命令格式：$pwd。
ls：显示目录内容，命令格式：$ls [options] [names]，选项可合用。
cd：改变工作目录，命令格式：$cd [directory]。

A3 文件创建、复制与删除

touch： 创建内容为空的文件，命令格式：$touch 文件名。
rm：删除文件或目录，命令格式：$rm [-r] [-f] [-i] file ...。
cp：复制文件或目录，命令格式：$cp [-i] [-r] file1 [file2...] target
-r。 如果 file 为目录，则 cp 将复制该目录及其所有文件。
mv：文件的搬移或更名，命令格式：$mv file1 target。

A4 文件属性

chmod：改变文件的读、写或执行权限，命令格式：
$chmod [who] operator [permission] file-list。
chown：改变文件的属主，命令格式：$chown [-R] [-h] owner file…。
chgrp：改变文件的组主，命令格式：$chgrp [-R] [-h] group file…。

A5 文件显示与连接

cat：用于显示文件内容或文件连接，命令格式：
$cat file1 file2 显示 file1 和 file2 的内容；
$cat file1 file2 > file3 将 file1 和 file2 合并成 file3。
more：显示文件的内容，命令格式：$more 文件名。
head：显示文件的前几行，命令格式为：$head [-n] [file…]。
tail：将文件从指定位置开始的内容全部显示到屏幕上：
$tail [+n] [lbc] file 从文件头加上 n 处开始显示；
$tail [-n] [lbc] file 从文件尾减去 n 处开始显示；
$tail –f file 间隔 1 秒循环显示文件新内容。
ln：建立指定文件的硬链接或符号链接，命令格式：
$ln [-s] [-f] [-n] file target。

A6 文件查找与比较

grep：查找字符串，命令格式：$grep pattern files。
find：从指定目录开始，递归地从子目录寻找匹配文件，命令格式：
$find dirname option-list。
diff：比较两个文本文件的差异，命令格式：diff [options] file1 file2。

A7 文件压缩与备份

compress：进行文件压缩，命令格式：$compress [-cfv] filename。
uncompress：解压缩文件，命令格式：$uncompress [-cfv] filename。
tar：用于建立磁带档案（文件系统的备份），或存到档案媒介或从档案媒介中读取文件，命令格式：$tar c|t|x [bvf] [tarfile] [bsize] [file_list]。

A8 输入输出重定向

<：输入改向，命令格式：$command < file。
>：输出改向，更新指定文件内容，命令格式：$command > file。
>>：输出改向，将执行结果接到指定文件内容后面，命令格式：
$command >> file。

附录 B 常用 vi 命令

B1 进入与退出 vi

进入 vi 命令格式：$ vi filename
vi 中的退出命令有以下几种：

:q	退出。当文件已被修改，vi 将在屏幕的底行显示提示信息。
:q!	强行退出。
:w	回写文件但不退出。
:wq	回写文件并退出。
:x	与 wq 相同，回写文件并退出。

B2 移动光标

↑或 k 键	把当前光标向上移动一行，保持光标的列位置。
↓或 j 键	把当前光标向下移动一行，保持光标的列位置。
→或 l 键	把当前光标向右移动一个字符。
←或 h 键	把当前光标向左移动一个字符。
$键	把当前光标移动到该行行末。
^键	把当前光标移动到该行行首。
w 键	把当前光标移动到该行的下一个字的首字符上。
b 键	把当前光标移动到该行的上一个字的首字符上。
e 键	把当前光标移动到该行的该字的末尾字符上。
^F	向前滚动一整屏正文。
^D	向下滚动半个屏正文。
^B	向后滚动一整屏正文。
^U	向上滚动半个屏正文。

在用 k、j、l、h 四个键时，可以在它们的前面加一个数字，这样在需要多次移动光标时不必多次按移动命令键。

B3 正文输入、删除、替换、恢复和查找命令

a	在光标的后面开始插入正文。
A	在光标所在行的行首插入正文。
I	在光标的前面开始插入正文。
I	在光标所在行的行末插入正文。
o	在光标所在行（当前行）的下一行的行首开始插入正文。
O	在光标所在行（当前行）的上一行的行首开始插入正文。
Esc	退出输入方式。
Backspace	输入方式下删除字符。
x	删除当前光标所在的字符。

nx	删除从当前光标开始的 n 个字符，n 为要删除的字符数。
dw	删除当前光标所在的字。
ndw	删除从当前光标开始的 n 个字，n 为要删除的字数。
dd	删除当前光标所在行。
ndd	删除从当前光标开始的 n 行，n 为要删除的行数。
rx	用 x 替代当前光标所在的字符。
nrx	用 x 替换 n 个字符，在替换完第 n 个字符后该命令自动停止。
u	废除最近的命令，恢复被修改或删除的内容。
U	把当前行恢复到修改它之前的状态。
/pattern	在缓冲区中向下查找指定的字符串 pattern。
?pattern	在缓冲区中向上查找指定的字符串 pattern。
n	重复上一次查找命令。
N	以相反的查找方向重复上一次查找命令。

B4 行编辑命令

键入“：”，并在屏幕底部的“：”号提示符下输入行编辑命令。

：set nu	显示正文的行号。
：set nonu	取消行号。
：1，$p	显示缓冲区的整个内容。
：r wqb	将文件 wqb 中的内容读入缓冲区，插入当前光标下。
：1，5w clh	将正文中 1 到 5 行的内容写到名为 clh 的文件中去。
：2，5d	将正文中 2 到 5 行删除。
：2，5t8	将正文中 2 到 5 行复制到第 8 行的后面。
：！ls	暂时转出 vi 编辑器，执行 shell 命令 ls。
：sh	暂时转出 vi 编辑器, 执行 shell 命令,键入^D 或 exit 返回正文。

附录 C GDB 常用命令

C1 启动 gdb

gdb	启动 gdb，不指定调试目标
gdb program	开始调试目标程序
gdb program [arglist] core	调试目标程序产生的 coredump 文件
gdb --help	获取 gdb 的命令行帮助信息

退出 gdb

quit

执行程序

run arglist	使用参数 arglist 启动程序执行
run	使用当前参数启动程序执行
set args arglist	为下一个 run 命令设置运行参数
set args	清空参数列表
show args	显示参数列表
show env	显示所有环境变量
show env var	显示指定环境变量
set env var string	设定环境变量
unset env var	从环境变量列表中删除指定变量

程序执行控制

continue	从程序停止处继续程序执行
step	单步执行程序直到下一行代码
si	单步执行程序直到下一条机器指令
next	执行下一行代码，包括函数调用
nexti	执行下一条机器指令
until [location]	运行直到一条指令（或指定地址 location）
finish	运行直到当前栈帧返回
return	跳过（不执行）选定栈帧

断点和观察点

break [file:] line	在[file]的第 line 行设置断点
break [file:] func	在[file]的 func 函数处设置断点
break *addr	地址 addr 处设置断点
break	在下一条指令设置断点

break … if expr	如果表达式 expr 非零，则断点设置有效
watch expr	为表达式 expr 设置观察点（当 exprt 值变化时停止程序执行）
info break	显示已设置的断点信息
info watch	显示已设置的观察点设置
clear	删除下一条指令处的断点
clear [file:] func	删除函数 func 入口处的断点
clear [file:] line	删除 line 行出的断点
delete [n]	删除第 n 个断点
disable [n]	禁用第 n 个断点
enable [n]	启用第 n 个断点
ignore n count	忽略第 n 个断点 cout 次

程序栈帧

backtrace	查看当前栈帧
frame n	选择第 n 个栈帧
up n	向上移动 n 个栈帧
down n	向下移动 n 个栈帧
info args	查看当前栈帧的参数列表
info locals	查看当前栈帧的局部变量

数据显示

print expr	显示表达式 expr 的值
x [/Nuf] expr	以指定格式查看内存地址 expr，参数定义如下：
N	指定显示的数据单元数量
u	单元大小：
	b, 单个字节
	h, 半字（2 个字节）
	w, 字（4 个字节）
	g, 8 个字节
f	数据显示格式：
	x，十六进制
	d，十进制
	u，无符号十进制
	o，八进制
	t，二进制
	a，对地址或相对地址
	c，字符
	f，浮点数